NRL Report 8517

# Validation of a Monte Carlo Simulation of Binary Time Series

R. E. JOHNSON

*U.S. Army Concepts Analysis Agency*
*Bethesda, Maryland*

H. L. WIENER AND D. ROQUE

*Systems Research Branch*
*Space Systems Division*

September 18, 1981

NAVAL RESEARCH LABORATORY
Washington, D.C.

| | | |
|---|---|---|
| **REPORT DOCUMENTATION PAGE** | | READ INSTRUCTIONS<br>BEFORE COMPLETING FORM |
| 1. REPORT NUMBER<br>NRL Report 8517 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>VALIDATION OF A MONTE CARLO SIMULATION OF BINARY TIME SERIES | | 5. TYPE OF REPORT & PERIOD COVERED<br>Final report on an NRL problem.<br>6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>R. E. Johnson*, H. L. Wiener and D. Roque | | 8. CONTRACT OR GRANT NUMBER(s) |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Naval Research Laboratory<br>Washington, DC 20375 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>61153N-14; RR-014-02-41;<br>NRL Problem 79-0719-0-1 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Department of the Navy<br>Office of Naval Research<br>Arlington, VA 22217 | | 12. REPORT DATE<br>September 18, 1981<br>13. NUMBER OF PAGES<br>18 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>UNCLASSIFIED<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

*Present address: U.S. Army Concepts Analysis Agency, Bethesda, MD 20014

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Monte Carlo evaluation
Simulation validation
Statistical tests
Binary time series

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report presents a statistical technique appropriate for the validation of a Monte Carlo simulation of a detection process whose results can be reduced to a finite time sequence of equally spaced events with dichotomous outcomes. The technique may be applied to other similar processes where there is not stationarity or steady state in the behavior of the process simulated. It is particularly useful in the case where multiple replications of the simulation have been obtained but only one observation of the real-world process is available. The Bahadur-Lazarsfeld representation of the probability distribution of the population consisting of all binary vectors having a given number of elements

(Continues)

20. Abstract (Continued)

is generated from the sample of such vectors produced by several independent replications of the Monte Carlo simulation. Then the validity of the simulation is determined to within a specified significance level, by comparison with a single realization of the actual process simulated. This validation technique admits correlation between events occurring at different times.

# CONTENTS

# VALIDATION OF A MONTE CARLO SIMULATION OF BINARY TIME SERIES

## INTRODUCTION

In this report a statistical technique is presented appropriate for the validation of a Monte Carlo simulation of processes such as detection processes, whose results can be reduced to a finite sequence of thresholding events having dichotomous, i.e., binary, outcomes. Moreover, the validation technique can be applied even when the process being simulated cannot be experimentally repeated. Thus the simulation's validity for the case examined can be determined, to within a specified level of statistical significance, with only a single observation of the real-world process being simulated. The statistical technique is nonparametric, it does not assume independence between events occurring at different times, and it does not require the assumption of any stationarity or steady state behavior of the process simulated.

The validation technique can be briefly described as follows. Each Monte Carlo replication of the simulation model produces a vector of $m$ binary elements. Based on this sample of binary vectors, a representation is obtained of the probability distribution of the population of binary vectors from which the sample was drawn. Using this representation the likelihood of occurrence of *any* vector of $m$ binary elements may be computed under the hypothesis that it comes from the same statistical population as the vectors generated by the simulation model. In particular the likelihood of the binary vector resulting from an observed run of the actual process under the same conditions that are represented in the simulation is computed. The question of the validity of the simulation model, at the significance level $\alpha$, is then resolved by observing whether the probability of the experimentally obtained vector exceeds the $\alpha$th percentile of the probabilities of the simulation-generated vectors.

The statistical technique described in this report constitutes a new application, to simulation model validation, of previous results concerning the representation of the probability distribution of dichotomous experimental responses. Included in this report is a description of how this technique was applied to the statistical validation of a particular simulation model used by the U.S. Navy to represent the surveillance performance of a system of undersea acoustic sensors.

## VALIDATION DEFINED

There is considerable diversity of opinion on what constitutes a validation of a simulation, and, for that matter, on how the term validation is defined. This report follows the currently accepted terminology, as in Fishman and Kiviat [1] and Steinhorst and Garratt [2], and defines *validation* as "testing the agreement between the behavior of the simulation model and the real system," as distinguished from *verification*, which is taken to mean "insuring that a simulation model behaves as the experimenter intends."

A multitude of different criteria and procedures have been proposed for the validation of simulation models. Some of these criteria are qualitative, such as the suggestion by Turing [3] that a model is valid if, given both a model's synthetic output and nonmodeling results in a similar

Manuscript submitted June 29, 1981.

format, an expert cannot discern which is the model result — or the widely used standby of visual comparison of model output with empirical data.

This report presents the view that a quantitative procedure for simulation model validation is preferable. A quantitative procedure is a statistical technique whose outcome is the determination whether the simulation model forecast does or does not agree with nonmodeling results, with a specified level of statistical assurance. A great variety of such quantitative procedures have been proposed for the validation of simulations. In the end, the choice of a statistical technique of validation must be governed by the structure of the model output data and the available real-world observations of the system being modeled.

## STRUCTURE OF THE VALIDATION DATA

The simulation model in question was designed to estimate the performance of an acoustic sensor in the detection of a target in the ocean. A principal output of the simulation is the set of instantaneous probabilities of target detection by the sensor at each hour throughout the duration of the target's specified maneuvers. The simulation requires a substantial collection of numerical data as input to the computer programs that make up the model. The U.S. Navy maintains a data base for use in modeling sensor performance, and this data base is periodically reviewed and approved. The model cannot be operated without input data. Therefore it is appropriate to treat the data base as fixed and to consider the combination of computer programs and data base as the simulation model to be validated.

The target's acoustic characteristics and the target track, i.e., the history of the geographical positions of the target at each hour, are provided as input to the simulation model. The model operates by replicating this track many times. During a particular replication, at each hour, the model decides that the sensor either is detecting or is not detecting the target, based on factors such as sensor-target range and geometry; the acoustic properties of the ocean in the sensor-target vicinity; sensor alertment due to possible detections at previous hours of this replication; and the magnitude of a Gauss-Markov fluctuation approximated by summing three independent Ehrenfest random walk terms (Feller [4]) at each hour. Thus for a target track lasting $m$ hours, a single replication by the model produces a vector of $m$ elements, each element being either 1 or 0, where 1 indicates the sensor is detecting and 0 not detecting the target at a particular hour. The instantaneous probability of detection at a given hour is then approximated by the mean over all replications of the vector element corresponding to that hour. The structure of these data is depicted in Fig. 1. The detection events occur at equally spaced intervals of one hour; however, the word "hour" could be replaced by the term "time step" wherever it occurs without affecting the accuracy of any statement. Moreover the events whose outcomes are represented by the binary vector elements need not be equally spaced over time.

The probability of detection of a given target by a given sensor is a strong function of the target's geographical position. This fact, in conjunction with the alertment effect of prior detections, makes it clear that a high degree of correlation exists between the detection events occurring at consecutive hours. Moreover, it would be most unusual for the same probability of detection to obtain at all points of a target track. Since a moving target is changing its position and aspect relative to the sensor as time progresses, no steady state will be reached in the finite duration of a target track.

Another factor affecting the selection of this particular validation technique is the lack of data concerning realizations of the simulated process. Multiple realizations of the same target track are practically impossible to obtain for the targets and sensors of interest. As is true of many processes that are the subjects of simulation, the expense and difficulty involved generally preclude making

| Elapsed Time (hours) | 1 | 2 | 3 | 4 | 5 | 6 | ... | 239 | 240 |
|---|---|---|---|---|---|---|---|---|---|
| Observed Detection History | 0 | 0 | 1 | 1 | 1 | 0 | ... | 1 | 1 |
| Model Replication | | | | | | | | | |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 1 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 | 1 |
| 3 | 0 | 0 | 1 | 1 | 0 | 0 | ... | 0 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| 50 | 0 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 |
| Predicted Detection Probability | 0.00 | .20 | .60 | .20 | .00 | .00 | ... | .00 | .60 |

Fig. 1 — Structure of validation data for an individual acoustic sensor. 240-h track, 50 model replications; 1: detecting the target; and 0: not detecting the target

repeated controlled experiments. It is impossible to achieve the degree of control necessary to obtain repeated real-world observations which would have identical inputs to the model. Detection histories for many different targets are available, but since the target tracks differ geographically, one cannot treat the detection histories of different targets as members of the same statistical population. Hence, observed samples of size one are the major constituent of the available records of the detection process simulated by the model.

The recorded history of detections of a given target by a given acoustic sensor is not considered to be a fixed standard to be matched by the simulation. The fluctuations in the physical processes involved in the origin, transmission, detection, and classification of an acoustic signal require the recorded history of gains and losses of contact with the target to be regarded as a sample of size one from a random population of unknown distribution. Hence, one is testing the validity of the simulation model as a representation of the statistical structure underlying the recorded history of detections. That is, the hypothesis being tested is that the unknown statistical distribution of which the observed history of detections constitutes a single sample point is the same as the unknown statistical distribution of which the model's replications constitute many sample points.

## STATISTICAL TECHNIQUE

As a particular target traverses the surveillance zone it generates a track history of detections for each sensor in the zone. For each sensor $i$ there is an observed vector $\hat{x}_i = (\hat{x}_{i1}, \ldots, \hat{x}_{im})$ of 0's and 1's from a probability distribution $p_o^i$ and the simulation generates $n$ vectors $x = (x_1, \ldots, x_m)$ of 0's and 1's from a probability distribution $p_i$. By use of the $n$ generated vectors, an estimate $\hat{p}_i$ is obtained of $p_i$. The simulated data are applied to $\hat{p}_i$ to obtain the sample distribution as an approximation to the population distribution. The test consists of determining whether the observed vector $\hat{x}_i$ has $\hat{p}_i$-value in the upper $1 - \alpha$ region of the sample distribution. If it does, the hypothesis that $\hat{x}_i$ comes from the distribution $p_i$ is accepted. The test is applied for all $i$ and many acceptances that $\hat{x}_i$ is from $p_i$ will confirm that $p_i$ is a good approximation to $p_o^i$, thereby validating the simulation. It is to be expected that due to statistical fluctuation some sensors will fail the test. Hence, a distinction must be made between validation of the simulation model in general and specific statements about the simulation of the individual sensors. Statements about the latter are simply understood to carry the uncertainty inherent in the statistical test itself.

The statistical hypothesis test uses the representation by Bahadur [5] and Lazarsfeld [6] for the probability distribution underlying binary sequences, which is summarized as follows. Assume the target track is $m$ hours long. Let $X$ be the set of all points $\hat{x} = (x_1, x_2, \ldots, x_m)$ with each $x_i = 0$ or 1, and suppose $p(\hat{x})$ is a probability distribution on the elements of $X$, that is, $p(\hat{x}) \geqslant 0$ for all $\hat{x} \epsilon X$ and $\Sigma_{\hat{x} \epsilon X} p(\hat{x}) = 1$. Let $E_p(\cdot)$ denote the expected value of the expression in parentheses when the distribution $p$ obtains. Then let

$$v_i = E_p(x_i) \qquad 0 < v_i < 1; i = 1, 2, \ldots, m; \tag{1}$$

and

$$z_i = (x_i - v_i)/\sqrt{v_i \cdot (1 - v_i)} \qquad i = 1, 2, \ldots, m. \tag{2}$$

Next define the family

$$\begin{aligned} r_{ij} &= E_p(z_i \cdot z_j) & i < j; \\ r_{ijk} &= E_p(z_i \cdot z_j \cdot z_k) & i < j < k; \\ &\vdots \\ r_{12\ldots m} &= E_p(z_1 \cdot z_2 \cdot \ldots \cdot z_m) . \end{aligned} \tag{3}$$

For $\hat{x} = (x_1, x_2, \ldots, x_m)$ define

$$p_{[1]}(\hat{x}) = \Pi_{i=1}^m v_i^{x_i} (1 - v_i)^{1-x_i} \tag{4}$$

and

$$\begin{aligned} f(\hat{x}) = 1 &+ \Sigma_{i<j} r_{ij} \cdot z_i \cdot z_j + \Sigma_{i<j<k} r_{ijk} \cdot z_i \cdot z_j \cdot z_k \\ &+ \ldots + r_{12\ldots m} \cdot z_1 \cdot z_2 \cdot \ldots \cdot z_m . \end{aligned} \tag{5}$$

Then for each $\hat{x}$ in $X$,

$$p(\hat{x}) = p_{[1]}(\hat{x}) f(\hat{x}) . \tag{6}$$

Thus $p_{[1]}(\hat{x})$ denotes the joint probability distribution of the $x_i$'s under an assumption that the $x_i$'s are independently distributed, and $f(\hat{x})$ represents the effects of correlation.

In this representation it is natural to refer to the parameters $r_{ij}$ as second order correlations, to the parameters $r_{ijk}$ as third order correlations, and so forth, culminating in $r_{12\ldots m}$, the $m$-th order correlation. The distribution $p$ then is said to have order $s$ if one correlation of order $s$ is non-zero and all correlations of order greater than $s$ are equal to zero. If a distribution is known to be of a certain order $s$, then the representation (5) need only extend to correlations of order $s$ or less.

The representation (5) and (6) was first given in [6] and was derived by induction on $m$. The proof shown here was given in [5] and is due to Bahadur.

PROPOSITION 1. For every $x = (x_1, \ldots, x_m)$ in $X$

$$p(x) = p_{[1]}(x) f(x) .$$

To establish proposition 1, let $V$ be the vector space of real-valued functions $g$ on $X$. Consider $V$ as an inner product space with inner product $(h,g)$ and norm $||g|| = (g,g)^{1/2}$, where the inner product is defined by

$$(h,g) = E_{p_{[1]}}(h \cdot g) = \sum_{x \epsilon X} h(x) \cdot g(x) \cdot p_{[1]}(x) .$$

It follows from (1) and (2) that the set

$$S = \left\{ 1; z_1, z_2, \ldots, z_m; z_1 z_2, z_1 z_3, \ldots, z_{m-1} z_m; z_1 z_2 z_3, \ldots,; \ldots; z_1 z_2 \ldots z_m \right\}$$

of functions on $X$ is orthonormal, i.e., $||g|| = 1$ for each $g$ in $S$, and $(h,g) = 0$ for $h$ and $g$ in $S$ with $h \neq g$. Since there are $2^m$ functions in $S$, since $V$ is $2^m$ dimensional, and since $p_{[1]} > 0$ for each $x$, the following proposition holds:

PROPOSITION 2. The set $S$ is a basis in the space of real-valued functions on $X$. This basis is orthonormal when $p_{[1]}$ obtains.

It follows, in particular, that each function $f$ on $X$ admits one and only one representation as a linear combination of functions in $S$:

$$f = \sum_{g \epsilon S} (f,g) \cdot g .$$

Now take $f = p/p_{[1]}$. Then

$$\begin{aligned} (f,g) &= \sum_{x \epsilon X} f \cdot g \cdot p_{[1]} \\ &= \sum_{x \epsilon X} g \cdot p \qquad\qquad (7) \\ &= E_p(g) \end{aligned}$$

for all $g$. Since $E_p(1) = 1$ and $E_p(z_i) = 0$ for $i = 1, \ldots, m$ by (1) and (2), it follows from the preceding paragraph and (7) that (5) holds, with the coefficients defined by (3). This establishes PROPOSITION 1.

In some applications the nature of the situation being studied or computational problems might make it necessary to assume a specified order to the distribution, even though the value of that order cannot be known precisely. If the selection is in error, then the "truncated" form of expression (5) will be in error and so will the resulting values of $f(x)$ and $p(x)$. As defined by (6), the estimated $p(x)$ may not be a probability distribution and may assume negative values for some $x$. This point is discussed by Bahadur [5]. In addition, to obtain the estimate $\hat{p}$ of (6) one must first obtain the estimate $\hat{f}$ of (5). The statistical fluctuation associated with the values $\hat{f}(x)$ is another source of error leading to negative values of $\hat{f}(x)$ for some $x$ and consequently for $\hat{p}(x)$. Because $\hat{p}$ may not be a probability distribution, values taken by $\hat{p}$ are referred to as likelihood values.

In the present application several considerations led to truncating the form of (5). A fourth-order approximation to the Bahadur-Lazarsfeld representation has been employed, truncating the expression in (5) after the fourth-order correlations and using a time window of 12 h, that is, assuming zero correlation between time steps more than 12 h apart. From observations of the detection process it seems reasonable that the correlation between instants separated by more than 12 h is negligible compared to the correlations between instants closer together, and the contribution of correlations of order greater than four is relatively insignificant. The truncation has also been necessary in order to keep the computer costs within reason.

The need has been established to develop a methodology using experimental evidence to estimate the proper size of the time window and the order of correlation. A Bayesian statistical method has been developed by Haskell [7] to be applied to detection data in estimating sensor performance values for various measures of effectiveness. The technique models the thresholding events as a stochastic process in continuous time with a well defined autocorrelation function. The autocorrelation function is characterized by one parameter which the technique estimates. Applying this method to the real-world observations of specific sensors whose performance is being simulated could provide a more precise assessment of the correlation order and length of the time window of the detection process.

## APPLICATION TO VALIDATION

The observed realization to be compared with the simulation results consists of a track history of duration $m$ hours together with the associated detection history. The simulation model is programmed to run with input parameters characterizing the observed situation. Then $n$ replications of the model are run, each producing a time series (vector) $\hat{x}_g = (x_{g1}, x_{g2}, \ldots, x_{gm})$, $g = 1, 2, \ldots, n$, of $m$ binary elements, where 1 denotes a detection and 0 no detection of the target by the acoustic sensor. Actual values of $m$ and $n$ used are $m = 240$ and $n = 50$. The $n$ binary vectors are used to estimate the parameters in the Bahadur-Lazarsfeld representation of the probability distribution corresponding to the population from which the $n$ sample vectors are generated. Simple unbiased estimators were chosen for all parameters. The estimators for the $v_i$'s are maximum likelihood when the distribution $p_{[1]}$ obtains, that is, when the $x_{gi}$'s are independently distributed. Since the $v_i$'s are assumed to be neither 0 nor 1, a reasonable correction is made should the data seem to indicate they are. The estimates are obtained by:

$$\tilde{V}_i = \begin{cases} 1/2n & \text{if} \quad \sum_{g=1}^{n} x_{gi} = 0 \\ 1 - (1/2n) & \text{if} \quad \sum_{g=1}^{n} x_{gi} = n \\ (1/n)\sum_{g=1}^{n} x_{gi} & \text{otherwise for } i = 1, 2, \ldots, m; \end{cases} \tag{8}$$

$$z_{gi} = (x_{gi} - \tilde{v}_i)\sqrt{\tilde{v}_i \cdot (1 - \tilde{v}_i)} \qquad i = 1, 2, \ldots, m, \quad g = 1, 2, \ldots, n; \tag{9}$$

and

$$\begin{aligned} \tilde{r}_{ij} &= (1/n)\sum_{g=1}^{n} z_{gi} \cdot z_{gj}, & 1 \leqslant i < j \leqslant m; \\ \tilde{r}_{ijk} &= (1/n)\sum_{g=1}^{n} z_{gi} \cdot z_{gj} \cdot z_{gk} & 1 \leqslant i < j < k \leqslant m; \end{aligned} \tag{10}$$

$$\tilde{r}_{12\ldots m} = (1/n)\sum_{g-1}^{n} z_{g1} \cdot z_{g2} \cdot \ldots \cdot gm.$$

The likelihood $p_g$ of the $g$-th model replication $x_g$ is given by

$$p_g = p(\hat{x}_g) = f(\hat{x}_g) \cdot \left[\prod_{i=1}^{m} \tilde{v}_i^{x_{gi}} \cdot (1 - \tilde{v}_i)^{1 - x_{gi}}\right], \qquad g = 1, 2, \ldots, n, \tag{11}$$

with $f(\hat{x}_g)$ as characterized by (5) and the $z$-values as given by (9). Then under the hypothesis that the random mechanism for the simulation is a model for the random mechanism underlying the observed sensor's detections, the recorded sequence of detections of that target by the specified sensor, tne binary vector $x = (x_1, x_2, \ldots, x_m)$, has likelihood $q = p(x)$ — relative to the Bahadur-Lazarsfeld representation of the $n$ model replications given by (6), using the $\tilde{v}$'s and $\tilde{r}$'s computed by (8) and (10) from the model replications. Once the numbers $q = p(\hat{x})$ and $\{p_g = p(\hat{x}_g) : g = 1, 2, \ldots, n\}$ have been obtained, it can be determined whether to accept the simulation model as valid at a significance level $\alpha$. The test procedure (a straightforward rank test) is to reject the hypothesis of association if the observed value $q$ falls below the $\alpha$-th percentile of computed values $p_g$. Define $N$ to be the number of elements in the set $\{p_g : p_g \leqslant q, 1 \leqslant g \leqslant n\}$. Then if $N \geqslant n\alpha$, the simulation model is determined to be valid in predicting the performance of the specified acoustic sensor in detecting that target. The model is rejected as not valid at the $\alpha$ significance level if $N < n\alpha$. Because the range of likelihood values spans several orders of magnitude, plots of the cumulative distribution are generally done in the form "log likelihood vs cumulative probability." The value of $N$ is obtained by first arranging the sequence log $(p_g)$, $g = 1, \ldots, n$, in ascending order, comparing each member with log $(q)$, and updating $N$ until the first value is found that exceeds log $(q)$. Suppose, for example, $\alpha = 0.10$ is specified. Then for $n = 50$, the criterion says that for the simulation to be accepted as a reasonable model, the number $N$ (the number of replications whose likelihood is no greater than $q$) must be at least 5. The reason a one-sided test is required here is that the higher the likelihood of the recorded detection history relative to the Bahadur-Lazarsfeld representation of the model replications, the better the agreement is between the recorded detection history and the simulation output. Hence one need only be concerned with rejecting the simulation model if the likelihood of the recorded detection sequence is low relative to the likelihoods of the model replications.

As mentioned before, the approximation to (5) can be negative for some vectors $\hat{x}$,in which case the hypothesis test cannot always be decided. Fortunately such occurrences turned out to be very rare. An alternate statistical technique was developed to deal with the very few instances in which it happened. An additional alternative to deal with this problem would have been to add correlation terms of higher order until the Bahadur-Lazarsfeld representation yielded a positive likelihood (if at all). This, however, is not guaranteed to work since all parameters are estimated. It was deemed too expensive to try for the very little reward to be obtained in incorporating just a few more cases. The simulation validation results are derived from the overwhelming majority of the cases in which the technique presented here was applied successfully.

Figures 2 and 3 illustrate an application of the Bahadur-Lazarsfeld representation. The sample log likelihood vs cumulative probability distribution curve is drawn using the likelihoods assigned by the Bahadur-Lazarsfeld representation to the model replications. The critical region is determined from the likelihood at which the cumulative probability curve reaches the tenth percentile ($\alpha = 0.1$). With $n = 50$ replications, if the replications are numbered in order of increasing likelihood the boundary of the critical region is the likelihood of the fifth replication. In Fig. 2 the log likelihood, -62.7, of the reported sequence of detections of that target by that acoustic sensor exceeds the log likelihood, -67.9, determining the critical region, so the simulation predictions are accepted as a good fit to the recorded observations. In Fig. 3 the likelihood of the reported sequence is less than the likelihood that bounds the critical region, so the simulation model in that case is rejected as not valid.

Users of the simulation model commonly run 50 replications of a target track; but before conducting a validation it was necessary to determine whether 50 model replications constitute a large enough sample with which to perform the validation hypothesis tests. Figure 4 shows the distribution of likelihoods (relative to their respective Bahadur representations) of three independent sets of 50 replications for the same target and acoustic sensor. This case seemed to exhibit more variability than usual between replications, yet the Smirnov two-sided test for goodness-of-fit (at the 0.20 significance level) indicated that the three samples could be assumed to have come from the same population. Thus, it was concluded that 50 replications of the model are sufficient for validation purposes. On the other hand, Fig. 5 is a similar illustration of the empirical distributions of three independent sets of 20 replications each for the same target and detecting sensor. Although these samples pass the Smirnov two-sided goodness-of-fit test at 0.10 significance level, at the 0.20

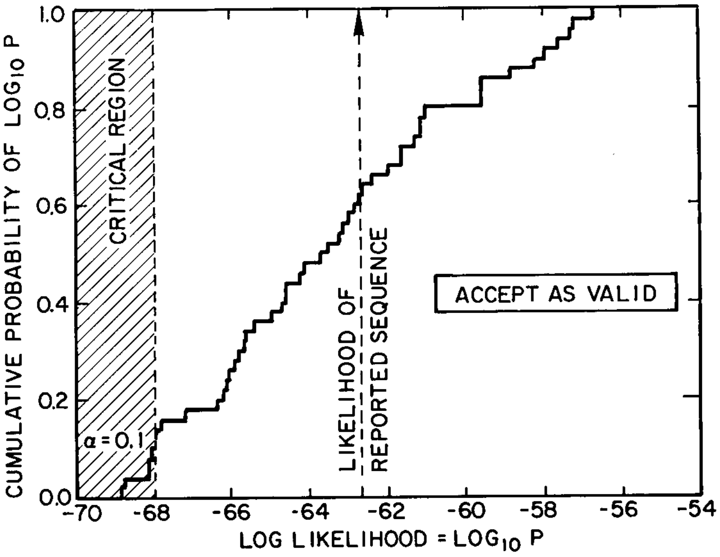

Fig. 2 — Distribution of 50 likelihood values good fit with observed data

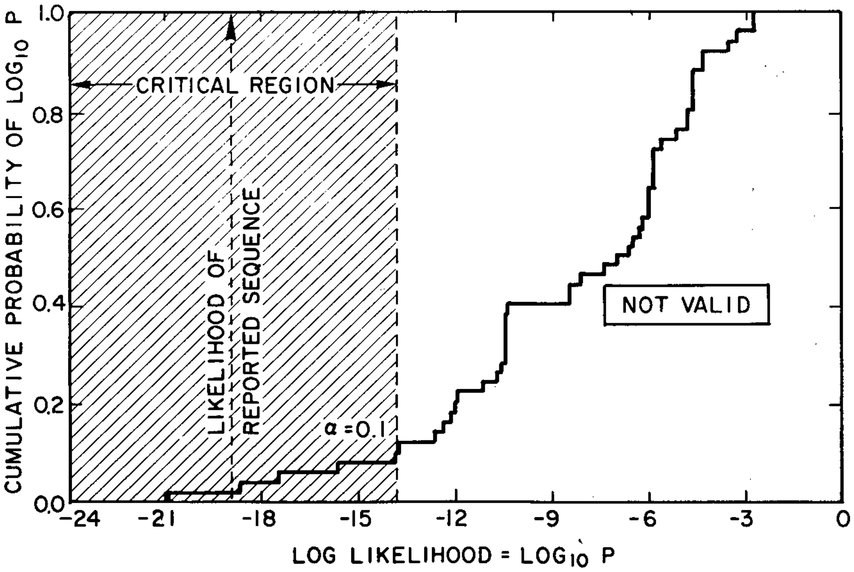

Fig. 3 — Distribution of 50 likelihood values poor fit with observed data

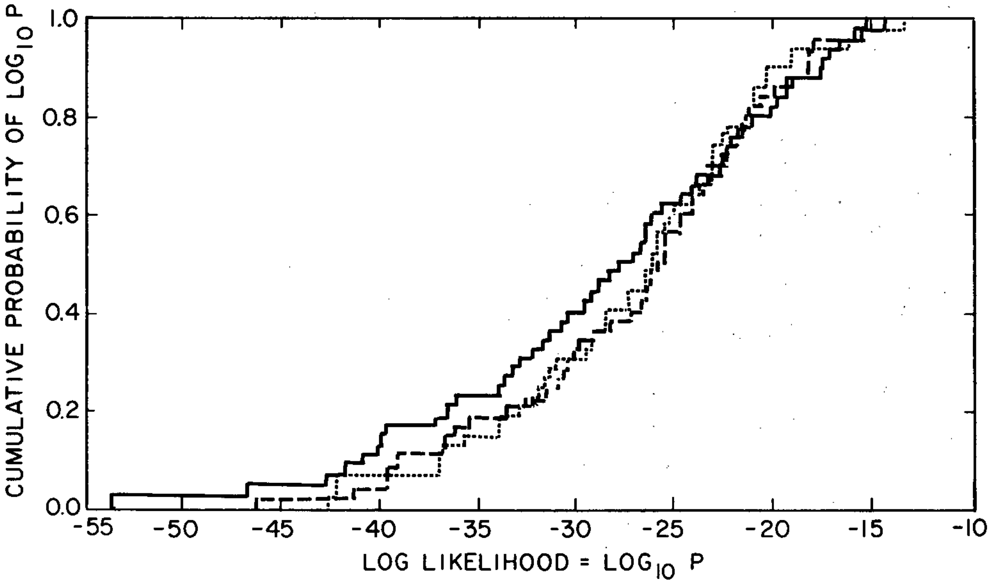

Fig. 4 — Smirnov test for three samples of size 50

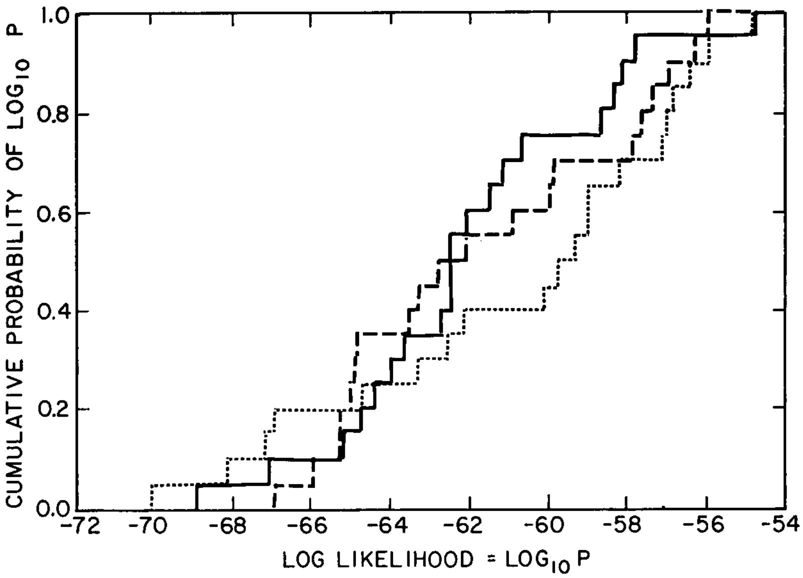

Fig. 5 — Smirnov test for three samples of size 20

significance level the test indicates the three samples did not come from the same population. Thus, it was concluded that the variability between samples of only 20 replications was too great to permit their use in a validation.

## VALIDATION RESULTS

The technique just described was applied to determine the validity of the simulation model in representing the detection performance of several different acoustic sensors. To be acceptable as valid, a model such as this should be able to represent any of a wide range of situations when given the characterizing parameters. Thus the true validation test lies in determining how well the model performs over a number of cases.

The tracks of nine different targets, whose detection histories were available, were simulated by the model, and 50 Monte Carlo replications of each target track were produced. The detection performance against these nine targets by the set of acoustic sensors that were operating was predicted by the model. The validation test was applied to the 50 detection performance vectors produced by the simulation for each sensor, along with the record of actual detections by each sensor. The results of these validation tests are summarized in Table 1. In this table a "+" symbol indicates those cases in which no detections of a particular target by a particular sensor were recorded. In all such cases the simulation model predicted a low enough level of detections to be accepted as valid. The symbol "*" denotes those cases where the model forecast was a good fit to the recorded history of detections and where there were detections of that target by that sensor recorded. The symbol "0" denotes those cases in which the simulation model output was found to be an inadequate fit to the observed record of detections. Table 1 shows that in the great majority of sensor-target combinations the predictions by the simulation model are a good fit to the recorded observations, at the 0.10 significance level. But one can identify certain sensors, particularly C, D, M, and R, whose performance is estimated inadequately by the model. The representation of these sensors within the data base of the simulation model has been singled out for investigation and possible improvement.

Table 1 — Validation Results

| Target / Sensor | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | * | 0 | + | + | + | + | + | 0 | + |
| B | + | + | + | + | + | + | + | 0 | + |
| C | + | + | 0 | * | + | 0 | + | 0 | + |
| D | + | + | 0 | * | + | 0 | * | 0 | * |
| E | + | + | + | + | 0 | + | * | * | + |
| F | + | * | + | + | + | + | + | + | + |
| G | * | + | + | + | + | + | + | + | + |
| H | * | * | + | + | + | + | + | + | + |
| I | * | 0 | + | + | + | + | + | + | 0 |
| J | 0 | * | + | + | * | + | + | + | + |
| K | * | 0 | + | + | + | + | + | + | + |
| L | * | * | 0 | + | 0 | + | * | * | + |
| M | * | 0 | 0 | + | * | * | * | 0 | * |
| N | * | 0 | + | + | + | + | + | + | + |
| O | + | + | + | + | + | + | + | * | + |
| P | * | 0 | + | + | * | + | * | 0 | 0 |
| Q | + | + | + | * | + | + | + | + | + |
| R | + | + | + | 0 | + | 0 | + | 0 | + |
| S | + | + | + | * | + | * | + | 0 | + |
| System | * | 0 | * | * | * | 0 | * | * | 0 |

*:Predictions accepted at 0.10 significance level. Detections recorded.
0:Predictions rejected at 0.10 significance level.
+:Predictions accepted at 0.10 level, but no detections recorded by this sensor.

These results suggest the following hypothesis about the model. It is to be recalled that the model was described as comprising both a computer routine and a set of numbers that characterize the situations, the sensors, and the targets. In a large family of cases the hypothesis of a good fit between model results and observed results was accepted; this would tend to indicate both a reasonably good family of mathematical routines and good supporting numbers in the model. On the other hand, in other cases the combination of the same mathematical routines with other supporting values produced results for which the hypothesis of a good fit was rejected. Consequently it seems worthwhile to postulate, as a hypothesis for further investigation, that the mathematical foundations of the simulation model are reasonable (or "valid"), but that the values characterizing some of the sensors are in error.

It is also worth noting that certain targets, such as number 2 and number 8, show a relatively high rate of rejection of the goodness of fit hypothesis. It is possible that the values used in the model to characterize these targets were in error. However, eliminating these targets from the data base will not make the results from all the sensors acceptable.

In addition to examining the individual sensor performance in the model, the validation methodology can be used to validate the results for families of sensors. Note that for validation purposes the detection performance of an entire set of sensors can be treated in exactly the same way as those for an individual sensor simply by defining a "system detection" whenever at least one of the set of sensors is detecting the target. By use of this scheme validation results were obtained for the simulation model's predictions of system detection performance; these appear in the bottom row of Table 1. The number of invalid estimates of system detection performance again suggests that the simulation model with the present data base should not be considered completely valid. When the input data representing certain sensors in the model have been corrected or refined, the model should be reexamined for validity.

## PROPERTIES OF THE TEST

The properties of the statistical test applied to validate the simulation remain open for further investigation. Of immediate interest is whether specifying the $\alpha$-th percentile of the sample distribution does result in a probability of rejection equal to $\alpha$. In addition there are such concerns as to how one should deal with alternative hypotheses and what kind of power does the test exhibit in evaluating alternatives. Only partial answers have been obtained to some of these questions.

Under the assumption that the simulation model is valid, the statistical test for a specific sensor-track combination constitutes a Bernoulli trial with rejection probability $\alpha$ and probability of no rejection $1 - \alpha$. Here the sample distribution (of the $n$ replications) is used as an approximation to the theoretical distribution; hence the actual rejection probability may appear to be data-driven. However the repetitions are identically distributed. Repetitions of the test itself under identical conditions should generate a proportion of rejections approximating the probability of rejection as the number of repetitions increase. In general there are $n + 1$ vectors each of length $m$. Ideally the probability of rejection should equal $\alpha$ when the $n + 1^{st}$ vector (sample vector) is from the same distribution as the first $n$ and it should equal the power of the test when the $n + 1^{st}$ sample is not from the same distribution. To make inferences about the probability of rejection, successive repetitions of the same test may be generated. The repetitions should be grouped into subsets of equal size from which a proportion of rejections may be computed for each one. The computed proportions are a sequence of iid. random variables. Taking a large enough number of subsets one can then appeal to the Central Limit Theorem and use classical statistical techniques to make inferences on the probability of rejection that results when applying the given test at a specific level $\alpha$.

A preliminary evaluation of the properties of the test has been conducted with the use of the computer. Sets of $n + 1$ random binary vectors were repeatedly generated from a known distribution of correlation order 1 and the test procedure was applied with a specified level $\alpha$ to determine whether the $n + 1^{st}$ vector did or did not belong to the population from which the first $n$ vectors came. The evaluation was conducted for various values of $m$ and various values of $n$ for each $m$. For the cases considered, the results appear to indicate that the probability of rejection is actually less than $\alpha$ for $m = 2$ or 3, but it approaches $\alpha$ from below as $m$ increases and is practically $\alpha$ for $m = 5$ and 6. This conclusion however cannot be accepted as general because the analysis was limited to a few known distributions of very simple structure. In a similar fashion the $n + 1^{st}$ vector was then generated from a known distribution of order 1 other than the distribution from which the first $n$ vectors were generated. The same procedure was applied, where the proportion of rejections now estimates the power of the test against a known alternative. Again, the analysis was limited but the test technique appears to discriminate very well.

An interesting problem revealed in the evaluation is that vectors with negative likelihoods arise as a function of the fluctuation of $f(x)$ around its theoretical value. In the cases considered the theoretical value is $f(x) = 1$. As $n$ increases the distribution of the $n$ values $\hat{f}(x)$ tends towards a spike at 1 and the number of vectors resulting in negative likelihoods decreases or disappears. In the future more formal attention should be paid to this problem since it is relevant in assessing the adequacy of the Bahadur-Lazarsfeld representation for practical application purposes.

Of a more general nature is one particular data structure that shows that the probability of rejection approaches $\alpha$ as $m$, the number of columns increase. The analysis consists of letting $n = 2^m$ where all binary vector columns consist of independent, identically distributed Bernoulli random variables with the probability of a 1 given by $v$. There are $(2^m)^{2^m}$ possible outcomes from which the Bahadur-Lazarsfeld representations may be obtained. To each of these outcomes one may associate $2^m$ possible $n$ + 1st or observed vectors yielding a total of $(2^m)^{2^m+1}$ elementary outcomes to be considered. First it is assumed that the observed vectors come from the same distribution as the sets of first $n$ vectors. Specifying $\alpha$, evaluating $N$ (as defined in application to validation) for each outcome and applying the rules of the test one may establish an association between $\alpha$ and the actual probability of rejection. Table 2 lists all possible outcomes when performing the test for the case $m = 1$. In this case the probability of rejection is independent of $\alpha$. For any $\alpha \epsilon (0,1]$ the probability of rejection is given by $p_r(v) = v(1 - v)$ depending only on $v$. The function $p_r(v)$ has domain $[0,1]$, is concave, and is symmetric around the point $v = 1/2$ where it achieves its maximum of 1/4. If it is now assumed that the observed vector comes from a different distribution, say a Bernoulli random variable with the probability of a 1 given by $t$, then the power of the test is also independent of $\alpha$ and given by $1 - \beta = v^2(1 - t) + (1 - v)^2 t$. The power of the test depends only on the relative values of $v$ and $t$. For larger values of $m$ the number of elementary outcomes to consider increases very rapidly, yet one can try to discern a pattern from the cases $m = 1$, 2 and 3. This is best illustrated by considering what happens when $m = 2$. The values of $N$ are either 0, 1, 2 or 4. This means that for values of $\alpha$ in the intervals $I_1 = (0, 1/4]$, $I_2 = (1/4, 1/2]$ and $I_3 = (1/2, 1]$ there correspond three different probability of rejection functions $p_{r1}(v)$, $p_{r2}(v)$, and $p_{r3}(v)$, depending only on $v$. These functions have range $[0, h_1]$, $[0, h_2]$ and $[0, h_3]$ respectively where $h_1 \epsilon I_1$, $h_2 \epsilon I_2$, and $h_3 \epsilon I_3$. The functions $p_{r2}(v)$ and $p_{r3}(v)$ corresponding to the larger values of $\alpha$ are concave and symmetric around $v = 1/2$ where they achieve their maxima $h_2$ and $h_3$. As $\alpha$ gets smaller, in this case $\alpha \epsilon I_1$, the function $p_{r1}(v)$ begins to behave in a different manner. It remains symmetric around $v = 1/2$ but in this instance it becomes bimodal. It jumps to a quicker maximum achieved at about $v = 0.3$ (also $v = 0.7$) and remains fairly close to its maximum for values of $v$ between $v = 0.3$ and $v = 0.7$. For values of $m > 1$ the number of non-overlapping subintervals covering $(0, 1]$ where $\alpha$ may be specified is given by $2^m - 1$. The rightmost subinterval always has length $1/2^{m-1}$. All others have length $1/2^m$. The length of the subintervals decreases rapidly and to each one corresponds a unique probability of rejection function $p_r(v)$ with range $[0, h]$ where $h$ is contained in the subinterval where $\alpha$ assumes its value. The functions $p_r(v)$ depend only on $v$ and are symmetric around $v = 1/2$. One can only speculate as to the general pattern followed by the functions $P_r(v)$ as $m$ increases. As $m$ increases $h \rightarrow \alpha$ and it appears that for smaller $\alpha$'s the functions $p_r(v)$ tend to achieve their maximum rapidly and remain close to this maximum for values of $v$ between the modes perhaps approximating a rectangular shape. This behavior supports that observed from the computer evaluation. If this is the case, as $m$ gets large and $\alpha$ gets small the probability of rejection approaches $\alpha$. The functions $p_r(v)$ for the case $m = 2$ are plotted in Fig. 6.

Table 2 — Test Outcomes for Case $m = 1$

| $x_g$ | | $\hat{v}$ | $p_g(x)$ | $q$ | | $N$ | |
|---|---|---|---|---|---|---|---|
| $g = 1$ | $g = 2$ | | $g = 1$<br>$g = 2$ | $x_o = 0$ | $x_o = 1$ | $x_o = 0$ | $x_o = 1$ |
| 1 | 1 | 3/4 | 3/4 | 1/4 | 3/4 | 0 | 2 |
| 1 | 0 | 1/2 | 1/2 | 1/2 | 1/2 | 2 | 2 |
| 0 | 1 | 1/2 | 1/2 | 1/2 | 1/2 | 2 | 2 |
| 0 | 0 | 1/4 | 3/4 | 3/4 | 1/4 | 2 | 0 |

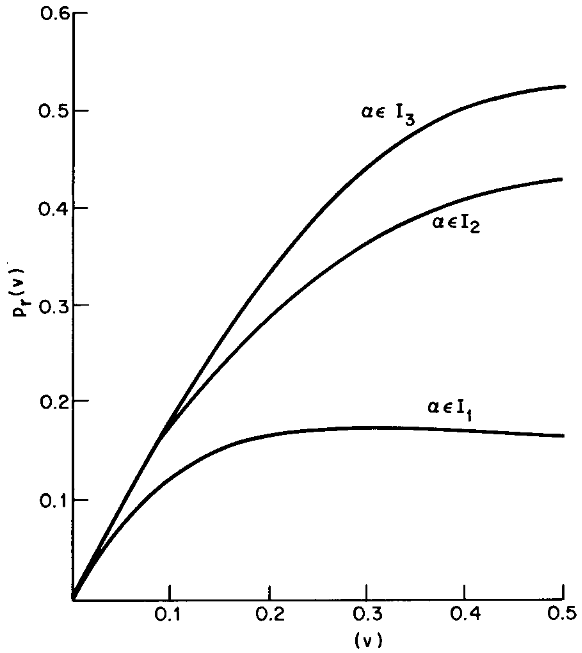

Fig. 6 — Functions $p_r(v)$ for the case $m = 2$

The analysis is by no means complete. So far only special cases with relatively simple structures have yielded some information on the nature of the statistical test. It is, however, a challenging problem that remains open for further consideration.

## SUMMARY

In summary, a method has been developed for the statistical testing of a Monte Carlo simulation whose output can be reduced to time series of binary data. This method has been implemented on a high-speed digital computer and has been successfully applied to determine that a particular simulation model, in combination with its approved data base, should not be considered valid. The validation technique presented here is applicable to simulations in a wide range of subjects, especially sonar systems, radar systems, other detection processes, and other processes which involve threshold crossing criteria to establish binary ("yes" or "no") outputs.

## ACKNOWLEDGMENTS

We would like to thank Robert S. Flum, Sr., and the referees for helpful comments and suggestions. The research in this report was supported by the Office of the Chief of Naval Operations, Strategic Analysis Support Group, and by the Office of Naval Research, under contract UF 118-99-100.

## REFERENCES

1. G. S. Fishman and P. J. Kiviat, "Digital Computer Simulation: Statistical Considerations," Report RM 5287-PA, Rand Corporation, Santa Monica, Calif., 1967.
2. R. K. Steinhorst and M. Garratt, "Validation of Deterministic Simulation Models," in "Proceedings of the Statistical Computing Section," pp. 1-10, American Statistical Assn., Washington, D.C., 1976.
3. A. M. Turing, "Computing Machinery and Intelligence," in *Computers and Thought* (Ed. J. Feldman and E. S. Feigenbaum), pp. 11-15, McGraw-Hill, New York, 1950.
4. W. Feller, *An Introduction to Probability Theory and Its Applications*, I, Wiley, New York, 1968.
5. R. R. Bahadur, "A Representation of the Joint Distribution of Responses to *n* Dichotomous Items," in *Studies in Item Analysis and Prediction* (Ed. H. Solomon), pp. 158-168, Stanford University Press, Palo Alto, Calif., 1961.
6. P. F. Lazarsfeld, "Some Observations on Dichotomous Systems" Sociology Department Report, Columbia University, New York, 1956.
7. R. D. Haskell "Correlated Opportunity Analysis of SOSUS Detection Data," Memorandum (CNA) 1543-75, Center for Naval Analyses, Arlington, Va., 1975.